IBM – Coursera

Data Science Specialization

Capstone project - Final report

Correlation between Number of crimes, Hardship Index and Surrounding features for Chicago

Gabriel Lima Novais – 2020

Table of content:

I. Introduction:

This report is for the final course of the Data Science Specialization. A 9-courses series created by IBM, hosted on Coursera platform. The problem and the analysis approach are left for the learner to decide, with a requirement of leveraging the Foursquare location data to explore or compare neighborhoods or cities of your choice or to come up with a problem that you can use the Foursquare location data to solve.

The main goal will be exploring the neighborhoods of Chicago city in order to extract the correlation between the Number of Crimes and its surrounding venues and Economic Index (Hardship Index of Community Areas)

The idea comes from the process of understanding why there are some community areas with more crimes than others, and two good features for that are Economic factors (such as unemployment, education and income) and geographic feature represented by surrounding venues. Surrounding places have peculiar characteristics such as shops, parks and sights that may be associated with the number of crimes

The target audience for this report are:

- Public Policy Makers.
- Researchers with interest in Econometrics and Sociology.
- Residents looking to understand the amount of crime in their city
- And of course, to this course's instructors and learners who will grade this project. Or to anyone who catch this shared on the social media showing that I can use Python data science tools.

II. Data description:

The main goal is taking the last 1000 crimes from Chicago and mapping all those crimes in community areas. Then, after finding latitude and longitude of community areas, insert some surrounding venues from Foursquare. With that data we want to understand how the characteristics and points of community areas influence in criminality levels.

The Source of Data:

- **Crime Data:**

  This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

  Link:
  https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2

- **Census Data:**

  This dataset contains a selection of six socioeconomic indicators of public health significance and a “hardship index,” by Chicago community area, for the years 2008 – 2012. The indicators are the percent of occupied housing units with more than one person per room (i.e., crowded housing); the percent of households living below the federal poverty level; the percent of persons in the labor force over the age of 16 years that are unemployed; the percent of persons over the age of 25 years without a high school diploma; the percent of the population under 18 or over 64 years of age (i.e., dependency); and per capita income. Indicators for Chicago as a whole are provided in the final row of the table.

Link:

https://data.cityofchicago.org/Health-Human-Services/Census-Data-Selected-socioeconomic-indicators-in-C/kn9c-c2s2

- **Coordinates:**

Current community area boundaries in Chicago. The data can be viewed on the *Chicago Data Portal* with a web browser. However, to view or use the files outside of a web browser, you will need to use compression software and special GIS software, such as ESRI ArcGIS (shapefile) or Google Earth (KML or KMZ), is required. But you can also have some data about these boundaries in github repositories which give us in a easier way Latitudes and Longitudes.

Link:

https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6

The cleaning process was done by after grouping all those data eliminating columns which were not needed. The result is in the figure below:

	Community Area Number	COMMUNITY AREA NAME_x	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX	Latitude	Longitude	Number of crimes
0	1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39	42.009120	-87.668648	18
1	2	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46	41.999316	-87.692394	9
2	3	Uptown	3.8	24.0	8.9	11.8	22.2	35787	20	41.966222	-87.658792	15
3	4	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17	41.968844	-87.685397	9
4	5	North Center	0.3	7.5	5.2	4.5	26.2	57123	6	41.950503	-87.681029	6

The data with Foursquare surrounding venues and with K-Means results can be illustrated in those figures below:

	CA	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	Clusters
0	Rogers Park	Mexican Restaurant	Accessories Store	Bus Station	Bakery	Grocery Store	BBQ Joint	Chinese Restaurant	Electronics Store	Falafel Restaurant	Farmers Market	Eye Doctor	3
1	West Ridge	Pizza Place	Discount Store	Mexican Restaurant	Seafood Restaurant	Eastern European Restaurant	Sandwich Place	Hotel	Design Studio	Cycle Studio	Farmers Market	Falafel Restaurant	3
2	Uptown	Chinese Restaurant	Park	Pizza Place	Seafood Restaurant	Mexican Restaurant	Storage Facility	English Restaurant	Electronics Store	Yoga Studio	Donut Shop	Ethiopian Restaurant	3
3	Lincoln Square	Pizza Place	Liquor Store	Bar	Automotive Shop	Yoga Studio	Eastern European Restaurant	Farmers Market	Falafel Restaurant	Eye Doctor	Ethiopian Restaurant	English Restaurant	3
4	North Center	BBQ Joint	Seafood Restaurant	Pharmacy	Dim Sum Restaurant	Video Store	Ethiopian Restaurant	Eye Doctor	Donut Shop	Farmers Market	Fast Food Restaurant	Field	3

After that the data finally become what we can see below:

	Community Area Number	CA	PERCENT OF HOUSING CROWDED	PERCENT HOUSEHOLDS BELOW POVERTY	PERCENT AGED 16+ UNEMPLOYED	PERCENT AGED 25+ WITHOUT HIGH SCHOOL DIPLOMA	PERCENT AGED UNDER 18 OR OVER 64	PER CAPITA INCOME	HARDSHIP INDEX	11th Most Common Venue	12th Most Common Venue	13th Most Common Venue	14th Most Common Venue	15th Most Common Venue	Clusters
0	1	Rogers Park	7.7	23.6	8.7	18.2	27.5	23939	39	Chinese Restaurant	Electronics Store	Falafel Restaurant	Farmers Market	Eye Doctor	3
1	2	West Ridge	7.8	17.2	8.8	20.8	38.5	23040	46	Hotel	Design Studio	Cycle Studio	Farmers Market	Falafel Restaurant	3
2	3	Uptown	3.8	24.0	8.9	11.8	22.2	35787	20	English Restaurant	Electronics Store	Yoga Studio	Donut Shop	Ethiopian Restaurant	3
3	4	Lincoln Square	3.4	10.9	8.2	13.4	25.5	37524	17	Farmers Market	Falafel Restaurant	Eye Doctor	Ethiopian Restaurant	English Restaurant	3
4	5	North Center	0.3	7.5	5.2	4.5	26.2	57123	6	Eye Doctor	Donut Shop	Farmers Market	Fast Food Restaurant	Field	3

III. Methodology:

The main goal will be exploring the neighborhoods of Chicago city in order to extract the correlation between the Number of Crimes and its surrounding venues and Economic Index (Hardship Index of Community Areas). A natural assumption is that with bad economic indicators and with geographic places such as parks and touristic points , for example, the number of crimes should be higher than others places

At the end, a regression model will be obtained. Along with a coefficients list which describes how each feature may be related to the number of crimes.

Indicators of economic hardship have been developed into an index to measure their economic conditions of Chicago Community Areas. This economic hardship index utilizes multiple indicators to provide a more comprehensive view of economic hardship than single indicators. Utilizing American Community Survey data, this fact sheet contains economic hardship index values for Chicago Community Areas. You can read more about them in:

https://greatcities.uic.edu/wp-content/uploads/2016/07/GCI-Hardship-Index-Fact-SheetV2.pdf

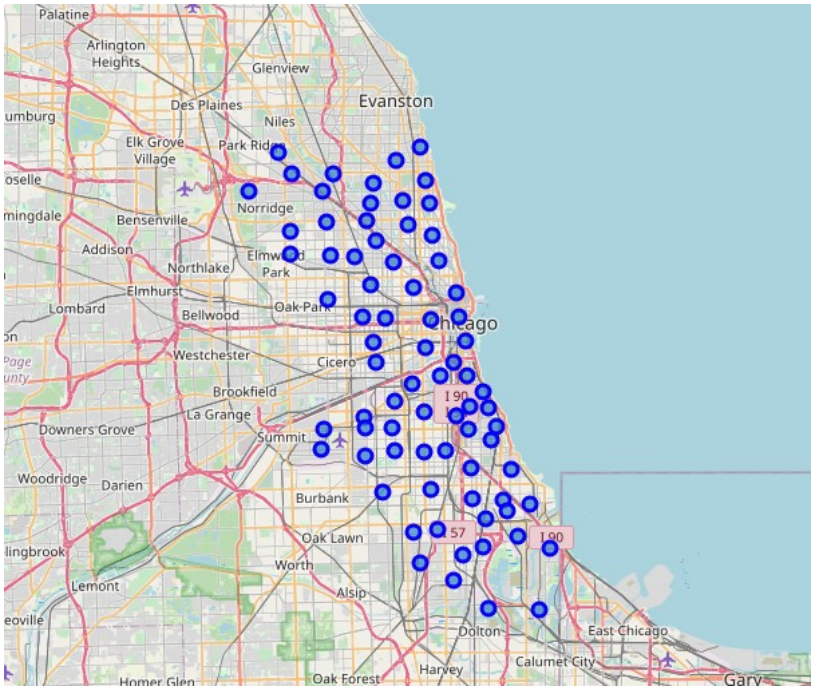
The clusters obtained from K-Means methods will give us a good variable that describes similar community areas based on surrounding venues. This will help us to understanding geographic features better.

Python data science tools will be used to help analyze the data.
Completed code can be found here:

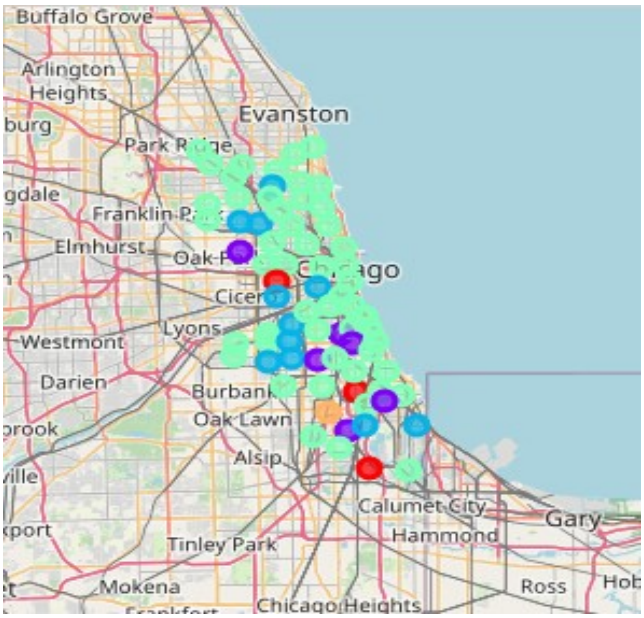
https://github.com/NovaisGabriel/Coursera_Capstone/blob/master/Coursera_Capstone.ipynb

1. K - Means:

In order to have a first insight of geographic features composed by the results of Foursquare for Chicago city no better way than visualization.



We can see from figure above that all points are uniformly distributed, and after the K-Means this map become:



2. Linear Regression:

Linear Regression was chosen because it is a simple technique. And by using Statsmodels library, implementing the model is quick and easy. Which is perfect to start the analyzing process.

The model will contain a list of coefficients corresponding to Hardship Index and Clusters types. R2 score (or Coefficient of determination) and Mean Squared Error (MSE) will be used to see how well the model fit the data.

IV. Results:

The results from the regression model can be seen in the figure below:

OLS Regression Results

Dep. Variable:	Number of crimes	R-squared (uncentered):	0.566
Model:	OLS	Adj. R-squared (uncentered):	0.554
Method:	Least Squares	F-statistic:	45.73
Date:	Thu, 30 Jul 2020	Prob (F-statistic):	1.97e-13
Time:	15:39:34	Log-Likelihood:	-278.41
No. Observations:	72	AIC:	560.8
Df Residuals:	70	BIC:	565.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
HARDSHIP INDEX	0.1539	0.037	4.199	0.000	0.081	0.227
Clusters	1.8560	0.793	2.341	0.022	0.275	3.437

Omnibus:	36.561	Durbin-Watson:	1.172
Prob(Omnibus):	0.000	Jarque-Bera (JB):	94.820
Skew:	1.645	Prob(JB):	2.57e-21
Kurtosis:	7.558	Cond. No.	33.5

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

We can then state that the coefficients are statistically significant, with a t-student above 2 and a p-value close to 0.05. Furthermore, we can see that the model has an approximate R2 of 56%, which indicates that there is a certain correlation between the independent and dependent variables. Another important factor to be said is that the higher the classification of the cluster, the greater the number of crimes, that is, clusters further away from the center are less likely to have high numbers of crimes.

V. Discussion:

The real challenge is constructing the dataset:

- Usually the needed data isn’t publicly available.
- When combining data from multiple sources, inconsistent can happen. And lots of efforts are required to check, research and change the data before merge.
- For data obtained through API calls, different results are returned with different set of parameters and different point of time. Multiple trial and error runs are required to get the optimal result.
- Even after the dataset has been constructed, lots of research and analysis are required to decide if the data should be kept as is or be transform by normalization or standardization.

It can be considered the most important process in the whole data science pipeline. Which can affect the most on the result.

On the other hand, choosing the suitable technique to construct the model is also a worthwhile process. As this report shows that, by applying a different method, the result can be improved.

VI. Conclusion:

We can get some meaningful and logical insights from the result. We can see that the model has an approximate R2 of 56%, which indicates that there is a certain correlation between the independent and dependent variables.

Doing this project helps practicing every topic in the specialization, and thus, equipping learners with Data Science methodology and tools using Python libraries.

Toward the person that went through this project, many thanks for the time and patient.

References:

https://greatcities.uic.edu/wp-content/uploads/2016/07/GCI-Hardship-Index-Fact-SheetV2.pdf

https://en.wikipedia.org/wiki/Community_areas_in_Chicago

https://data.cityofchicago.org/